AN INVESTIGATION INTO MACHINE RECOGNITION OF VOWEL-LIKE SOUNDS AND THEIR ALLOPHONES IN ONE SYLLABLE WORDS

John Paul Hydinger

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

An Investigation Into Machine Recognition
of
Vowel-like Sounds and Their Allophones
in
One Syllable Words

by

John Paul Hydinger and Frederick Michael Stubbs

Thesis Advisor: Robert C. Bolles

June 1972

An Investigation Into Machine Recognition of Vowel-like Sounds and Their Allophones in One Syllable Words

by

John Paul Hydinger
Lieutenant, United States Navy
B. S., United States Naval Academy, 1968

and

Frederick Michael Stubbs
Lieutenant, United States Navy
B. S., University of Utah, 1965

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
June, 1972

ABSTRACT

The goal was to recognize sustained vowel-like sounds and their allophones in one syllable words. A bank of filters and a digital sampler provided a data base for a polynomial curve fitting routine. The frequency range under investigation was 500-1000 Hz. A COMCOR CI 5000 analog computer and an XDS 9300 digital computer were used. Although coefficient correlation was ineffective, several recommendations for system improvement are made.

TABLE OF CONTENTS

I. INTRODUCTION

Attempts at speech recognition use either special purpose hardware or computers. In both cases, filter banks are often used. The majority of the work in the field has been formant and frequency analysis.

The goal was to achieve a recognition algorithm for sustained vowel-like sounds and their allophones in one syllable words. It was assumed that a voiced audio signal could be broken into eight frequency bands ranging from 500 to 1000 Hz and the respective audio curves fitted to polynomials. It was further assumed that similar curves have similar coefficients.

A hybrid system, consisting of a COMCOR CI 5000 analog computer and a Xerox Data Systems 9300 digital computer, is used to effect a speech recognizer. Figure 1 is a diagram of the system.

Two experiments were conducted prior to system implementation. Experimentation using various frequency ranges was attempted in order to resolve a frequency conflict. In the first experiment subjects listened to random words, whereas in the second experiment brush recordings of the same words were studied.

The heart of the analog system is a parallel bank of eight band-pass filters. Their output is smoothed, sampled, and sent to the XDS 9300 for analysis. Figure 2 is a diagram of the complete analog system.

A digital program performs a fifteenth degree polynomial fit on each of the eight audio curves that are sampled from the analog computer. The program then outputs eight sets of normalized coefficients for elementary analysis. System noise is eliminated digitally, and zero data

points occurring at the end of words are completely overlooked by the polynomial fitting routine. Several program modifications were incorporated and their results discussed.

II. BACKGROUND INFORMATION

A. TRADITIONAL APPROACHES

In the investigation of speech recognition by the direct analysis of a speech wave (Reddy, 1966), the goal was to produce a phonemic transcription of a connected utterance which is readable and bears a satisfactory resemblance to what was said. The problem was confined to a single cooperative speaker so that writing, adjusting and testing programs would be easier. It was felt that a "tune-in" process would adapt the program to a wider variety of speakers. No attempt was made to group the phonemes into words or higher level linguistic units.

The concepts which were considered, such as amplitude normalization and time normalization, show some insight. In the case of sustained sounds and one syllable words, though, time normalization may not be necessary. It does not seem realistic, however, that the "tune-in" process could overcome the lack of generality in the original program.

A procedure for segmenting connected speech (Reddy and Vicens, 1968) performs smoothing and differencing operations on the digitized acoustic waveform to generate parameters which are used to determine whether the characteristics of a sound are changing or similar. Parts that possess similar parameters are grouped together to form sustained segments, resulting in the segmentation of connected speech into parts approximately corresponding to phonemes.

Smoothing looks like a reasonable operation to perform on waveforms before they are compared. A question that arises, though, is whether the smoothing should be done in the analog circuit or after the information has been digitized. Perhaps, too, one smoothing operation is not enough.

A successful limited speech recognition system (Bobrow and Klatt, 1968) operates within limitations along a number of dimensions. Rather than use continuous speech in which segmentation is a problem, the approach is to work with messages with easily delimited beginning and termination points. The set of messages is limited in number; at any one time the vocabulary to be distinguished can contain up to about 100 items. However, an item need not be a single word, but may be any short phrase. The system is useable by any male speaker, but must first be trained by him. The system, LISPER, is not designed to work well simultaneously for a number of different speakers, or achieve good recognition scores for an unknown speaker. The training period consists of a period of closed loop operation in which the speaker says an input message, the system guesses what he says, and he responds with the correct message. The recognition algorithm is a program that learns to identify words by associating the outputs of various property extractors with them. Each property has a corresponding feature state which may imply that the property is irrelevant for the current time interval, the property is relevant but not present, or the property is both relevant and present.

Several advantages of this approach are:

1. A precise segmentation of the utterance is not required.
2. The utterance need not be a single word.
3. Features may be added to the system to provide desirable redundancy.
4. The feature approach permits the introduction and testing of linguistic hypotheses.

Two main disadvantages are:

1. The current implementation is not speaker independent.

2. The system will degrade in performance as the length of the vocabulary is increased or as the number of speakers that it can simultaneously recognize is increased.

The differential effects upon vowel intelligibility of various degrees of time compression and frequency division were examined both with and without time restoration (Daniloff, Shriner and Zemlin, 1968). A male speaker and a female speaker were used. For a given percentage of distortion, frequency division degrades vowel intelligibility more severely than time compression. Restoring time to normal for frequency-division speech does not enhance intelligibility. Vowel confusions under time compression are related to duration; those for frequency division conditions appear to be closely related to the perception of Vowel Formant Two, and to a lesser degree, Vowel Formant One. Patterns of male and female vowel confusions are generally much alike for all conditions and types of distortion. Results tentatively indicate superior female vowel intelligibility under all conditions of distortion, the advantage being largest for frequency division and somewhat less for time compression. These results suggest that over a limited range of frequency division up to forty percent, vowel phonemic quality is relatively unaffected by proportionate shifting of fundamental frequency and formant structure, indicating that a "relative-vowel" hypothesis of vowel phonemic quality may hold for limited shifts in the frequency of vowel spectra.

The idea that vowel phonemic quality may hold during normalization is extremely important. However, the statement that vowel confusions under time compression are related to duration conflicts with another study (Seo, 1968). The method yields time compressed speech which is of normal pitch, and highly intelligible. It utilizes a systematic approach in which portions of phonemes are sectioned out without destroying cognitive qualities.

Another process for the extraction of significant parameters of speech involves division of the speech spectrum into convenient frequency bands, and calculation of amplitude and zero-crossing parameters in each of these bands every ten milliseconds (Vicens, 1969). In the software implementation, a smoothing function divides the speech spectrum into two frequency bands (above and below 1000 Hz). In the hardware implementation, the spectrum is divided into three bands using bandpass filters (150-900 Hz, 900-2200 Hz, and 2200-5000 Hz).

As in many other approaches, considerable effort is spent investigating from one-fourth to one-half the range of human hearing. Although this may be the correct approach to take, the experiments discussed in the next section would seem to indicate otherwise.

In an interview at Stanford Research Institute (Walker, 1972) it was suggested that, rather than concentrate solely on sustained sound, it might be worthwhile to look at the dynamics of sounds. It was further suggested that the upper limit of the frequency range to be investigated be increased to 10 KHz.

An earlier conversation with some of the technical people at Pacific Telephone revealed that a frequency range of 500-1000 Hz would result in a highly intelligible sound to a human listener. If this is the case, either:

1. The intelligibility is context dependent.

2. A significant speech parameter is being overlooked by the people who are investigating the frequencies above 1000 Hz, feeling that such investigation is necessary to insure adequate information.

In particular, a considerable amount of time is spent looking for significant vowel information between 3000 and 4000 Hz. Section II will discuss this conflict in more detail.

B. GOAL

The initial goal was to attempt to program a hybrid system to recognize phonemes, or basic sustained sounds, with particular emphasis on the differences of similar sounds. The sustained sound, however, is static and therefore unrealistic in nature. The goal was then modified so that the investigation would include some sustained vowel-like sounds, then some one syllable words containing those sounds, and finally an attempt to break down the word to study the dynamics of the vowel-like sound.

C. PRELIMINARY ASSUMPTIONS

The original premise was that the voiced sound could be broken into different frequency ranges, and that a subroutine could be used that would perform a polynomial fit to each of the filtered audio signals. The coefficients from these fits would then be used as a data base for phoneme recognition. This implies that similar curves will have similar coefficients. A comparison of the coefficients from two sets of data that are supposed to represent the same sound leads to the theory that a unique correlation exists in some subset of those coefficients.

Correlation infers that some subset of coefficients of a sound is a multiple, plus or minus some error tolerance, of the same subset of coefficients of the same sound said at another time. This subset will be referred to from now on as the "characteristic subset" of a sound.

D. VOCABULARY

Any sound that is not a single vowel-like sound or a one syllable English word containing that vowel sound is outside the domain of discussion. A vowel-like sound excludes some vowel pronounciations, such as ī; it includes dipthongs such as ou in the word though. However, ou is excluded in words such as out.

E. SYSTEM OVERVIEW

There are three phases to speech recognition:

1. Manipulate and sample an analog signal.
2. Digitally analyze the samples obtained from the analog computer.
3. Apply a recognition algorithm to the results of the digital analysis.

In this research, an audio signal is filtered into eight pass bands after a comparator is keyed by the excitation voltage. The output from the filters is smoothed prior to the digital sampler. At the point of smoothing, the envelopes of the filtered signals may be looked at on the brush recorder. The digitized samples are passed to a software buffer in the digital program. After sampling is complete, program analysis attempts to fit the sample points with a high order polynomial.

Two of the three phases have been satisfied. The current state of the project does not use a recognition algorithm.

III. INITIAL EXPERIMENTATION

There was a contradiction between the information gathered at SRI about relevant frequency ranges and that obtained from Pacific Telephone. Consequently, experimentation was begun by wiring two Kronhite filters in series to create a band-pass filter with a variable range. After a microphone input and an earphone output were connected, the upper and lower bounds of the pass band were varied to determine the comprehensability of randomly selected words. Several sets of twenty-five random words were chosen to be read by three speakers, including one female speaker. The listener was to wear the headphones and write down each word as he heard it. Eight listeners were selected, given no background information, and asked to put on the headset, face away from the speaker, and write down whatever words they heard. By so doing, no visual aids to speech perception were available to the listener (i.e., lip movement). Furthermore, care was taken to ensure that the listener could not hear anything except what came through the headset.

The initial frequency range used was 500-1000 Hz as this was the range of primary interest. It was found that the comprehensability of the words that were selected ranged from a low of 17 out of 25 correct to a high of 19 out of 25 correct; the largest majority being centered at 18 out of 25 words. In 100% of the cases, the vowel sounds were totally perceptible. Also in every case, the sounds that were incorrectly transcribed were words beginning with th, d, f, and s, the sounds all sounding somewhat alike to every listener. The next step was to change the lower bound of the filter to zero in order to discover

any further information that might be available at the lower frequencies. In looking at the results of these tests, it was determined that no increase in information was gained. The conclusion was that the lower bound of 500 Hz was reasonable.

The next frequency range investigated was 1000-2000 Hz, with somewhat startling results, for there was almost a total loss of word recognition. This made the frequency range of 500-1000 Hz a necessary condition for speech recognition.

As a check on the primary upper limit of 1000 Hz., the range 500 to 2000 Hz. was investigated. This was done to establish an upper frequency bound on the remaining information. This proved to be sufficient as a one hundred percent comprehension from all listeners was obtained. To further narrow down this critical range, the upper limit of the band pass was lowered to 1500 Hz. It was found that the same level of understanding was present. This upper level was lowered to 1400 Hz. without any information loss, but below this level the same difficulties were encountered as in the primary frequency range (i.e., 500-1000 Hz.).

The preceding experiment brought to light a salient point: Human beings possess some other faculty for speech understanding besides just a complete frequency spectrum analysis. But there are obviously critical frequency ranges because all words could not be understood at frequencies outside the critical range.

It should also be noted that obtaining center frequencies for filters in the range around 1500 Hz. is very unreliable due to the inaccuracy of the hardware. This is so because the CI-5000 was designed to work efficiently only at frequencies below 1000 Hz.

At this stage of the experimentation the brush recorder indicated the original premise, 500-1000 Hz is both a necessary and a sufficient condition for speech recognition, was correct. Efforts were concentrated on looking at the words and sounds which were earlier confused by the listeners. After several recordings, the fact was established that there were differences between the difficult to discern words in the upper frequency ranges (800-1000 Hz).

Based on the results of the experiments, it would be reasonable to expect the primary frequency range to contain enough information to make speech recognition possible.

IV. ANALOG SYSTEM

The input to the analog system is a microphone, the audio output of which goes through a pre-amplifier and from there is fed, via the key. keying circuit, to a bank of eight paralleled band-pass filters. The output of each filter is connected to a smoothing circuit, and from there to the channels of the digital sampler, which in turn feeds data to the digital computer (see figure 2).

A. COMPARATIVE NETWORK

The comparative network (see figure 3, part A.) acts as a keying circuit for the analog system. Its function is to start the analog data gathering when a person speaks into the microphone. This was necessary in order to minimize the timing problem of speech recognition.

The diagram shows two inputs to the comparator (C $\overline{00}$); one being the audio input signal and the other a reference signal. By adjusting the potentiometer (P), the exciting voltage level can be altered. It is normally set just above the noise level so that random noise will not accidentally key the circuit.

The output of the comparator is normally false or zero; when the circuit is keyed, even for an instant, delay flip flop zero (DFO) changes from false to true for a period of time determined by a dial setting. This in turn puts a true signal into T100 (TEST(1) in digital program) and interrupt 52 is enabled.

In order to control the system input, a digital three position switch (DSO) is employed. As long as the switch is in the middle or ground position, it acts as a short circuit and prevents T100 from going true.

When placed in either of the two true positions, it acts as an open circuit and T100 can be enabled. Thus, to key the system, DSO must be set to true and the speaker must then excite the circuit.

B. BAND-PASS FILTER

It was necessary to build eight band-pass filters on the CI-5000 analog computer. They had to be realizable component-wise. Most textbook filters were realizable, but impractical as eight could not be made with the existing hardware. The filter chosen was selected with reluctance for although it met the aforementioned requirements, it was a low Q or low resolution filter.

The diagram (figure 4) shows two amplifiers (A_1 and A_2), two integrators (I_1 and I_2) and three potentiometers (P_1 thru P_3). Potentiometer one controls the center frequency of the filter, while potentiometers two and three control the band width. Table one lists the actual components used and Table two lists both the associated potentiometer settings and the filter frequency ranges.

C. SMOOTHING CIRCUIT

A smoothing circuit was incorporated into the system, again, due to hardware limitations; this will be discussed in detail in the Digital Program Development section under smoothed data. The output of each of the filters is fed into a separate smoother and from there to separate channels of the digital sampler. The function of the circuit is to trace the envelope of the audio curve.

D. SAMPLING FREQUENCY

The sampling frequency is controlled by two things; first, the frequency generator used and second, the frequency divider (PSET CTR) (see figure 3, part B.). In order to attain a sample frequency of 200 samples per second a 10 K c frequency generator is used in conjunction with a division by 50, set into the PSET CTR. This generates a pulse into delay flip flop one (DF1) every five milliseconds. DF1, in turn, enables interrupt 52 for .1 millisecond during which time a sample is taken by the eight used channels of the digital sampler simultaneously.

V. DIGITAL PROGRAM DEVELOPMENT

The output from the CI 5000 is transferred to the XDS 9300 by means of a hardware link between the two machines. When an interrupt occurs, control is transferred to the subroutine which handles the buffer indexing, and which also calls the system subroutine which loads the buffer. The digitized samples from the analog computer are stored in the buffer until the complete set of data has been gathered. Once this has occurred the interrupt is disabled and the analysis begins.

An orthogonal least-squares curve-fitting technique is applied to the data from each of the eight filters, and the resulting polynomial coefficients are printed. The coefficients are used to compute values for the dependent variable, which is currently plotted by hand to compare to brush recordings of the same data.

A. INPUT DATA AVERAGING

Due to core limitations, which will be discussed in the following section, there was not sufficient space to store all of the samples taken if the sampling frequency was high (i.e., around 1000 Hz). Therefore, an averaging technique was employed. What actually occurred was simply a temporary buffering of a summation of several consecutive points before their inocrporation into the data set to be used by the curve-fitting routine. From two to ten points were averaged at various times. This technique was later found to be unnecessary and too costly timewise, and was therefore eliminated.

B. CHANGING SAMPLING FREQUENCY

An initial, but mistaken, assumption was that samples could be taken up to and including one sample every millisecond on each filter. Thus, for each channel one thousand data points could theoretically be obtained over a period of one second. However, due to limitations of core storage, a maximum sample size of 500 data points per filter became the upper limit. This limit could have been extended by the use of overlaying techniques in the XDS 9300 memory, but these techniques were found to be too slow to effectively take data at higher rates. The data that were obtained in using sample frequencies up to 500 samples per second had large discrepancies. There was an even more severe limitation in the sampling frequency in that samples could not be taken any faster than 200 points per second; thus, one sample every five milliseconds. The problem that existed at higher frequencies was that the buffering subroutine was too slow, causing a stacking of analog interrupts and resulting in lost data points.

Now that an upper bound had been established for both the sample frequency and the sample size, samples could be taken over a total time interval of two and one half seconds. However, because of the nature of the previously defined vocabulary, samples need only be taken for one second or less, with the mainstream of words lasting only one-half to three-quarters of a second. It was for this reason that the sample data set normally consisted of one hundred or one hundred and fifty data points representing one-half or three-quarters of a second respectively.

C. RAW VERSUS SMOOTHED DATA

Early in the research, the data was being fed directly from the hardware filters to the digital sampler, the resulting data being termed "raw data." In attempting to look at the representative plots on the brush recorder, it was discovered that the frequency of the filtered audio signal was too high for the brush recorder's mechanical recording arm to follow accurately. In order to alleviate this problem, a smoothing circuit was constructed external to the analog computer (figure 5). The function of this circuit was to smooth the data in such a way as to present the envelope of the original high-frequency curve. The plotting of this curve was within the mechanical capability of the brush recorder, and in fact led to the next step in data manipulation. For it was this smoothed curve that was, in fact, interesting. Therefore, instead of the data being fed directly from the analog filters to the digital sampler, the signal was smoothed first (see figure 2).

Sampling the higher frequency curve often gave misrepresentative data, whereas sampling the envelope resulted in much more consistent data. The curve obtained by sampling the raw data was found to be dependent upon two factors: (1) the initial point of sampling; and (2) the sampling frequency used. This was not the case when sampling on the envelope of the curve, for it was immaterial where the sampling started or what the interval was; the curve remained almost the same using recorded input.

D. NOISE

As was just mentioned, the curves that came from recorded input were almost the same. It was this fact that led to the assumption that random noise was present in the system. The primary question was just how extensively the noise affected the input data. A way to determine this was to reduce the keying bias to zero, thereby causing the analog program to take data without an exciting voltage. Thus, the only data taken would be noise in the system.

After several data runs of this type, the magnitude of the noise was found to be approximately one one-thousandth that of the desired input. It was therefore decided to truncate all information that was contained at the noise level and retain only three significant digits from the direct analog input. To ensure that the method was successful the initial testing process used in finding the noise was rerun. With a zero input to the system, all data was successfully truncated to zero. Furthermore, identical inputs produced more nearly identical outputs.

E. NORMALIZATION

In attempting to compare two sets of coefficients, it was noticed that there was often a correlation if a scaling factor was applied to one of the sets of coefficients. The difference in the size of the coefficients was possibly due to the change in volume when saying a word from trial to trial. Consequently, the coefficients would differ from trial to trial. Thus, an attempt was made to normalize the equations based on the setting of the high order coefficient to a particular constant thereby causing the other coefficients to be scaled.

This technique gave very promising results for discrete sets of trials, but when the intersection of the characteristic subsets was taken, the resulting subset was found to be empty, as no correlation could be obtained for all data. One of the interesting points that this particular method reinforced was the fact that it was much easier to attempt correlation with a single speaker than to attempt correlation between different speakers.

It is important to note that the aforementioned normalization is only amplitude normalization. The concept of time normalization has not been employed, because its importance has been realized only in the most recent stages of research. The idea of time normalization will be treated later in the paper.

F. VARIABLE WEIGHTING FUNCTION

Initially it was felt that unweighted data would suffice in the analysis of a filtered signal. The reasoning was that if sounds could be distinguished visually on the brush recorder, then fixed time sampling using a ten millisecond time interval would yield satisfactory results.

Consideration was then given to the idea of equating the weight given to a particular data point to the value of the data point. The intent of this was to emphasize the larger peaks and deemphasize the smaller peaks. By so doing, the curve fitting routine would place greater weight on the peaks when calculating coefficients. This was also intended to give a zero weight to data points with zero value.

If the sound being analyzed does not cover the full time interval that is being sampled, then zero data points appear at the end of the data set. This causes the curve fitting routine to attempt to fit not

only the non-zero data points, but the zero data points along the x-axis as well. By requiring the polynomial to fit the x-axis, it was believed that less accurate results would be produced than if the fit were restricted just to the non-zero data points. The problem was alleviated by setting the weights of the zero data points equal to zero.

The equating of weights to values neglected the possibility that a small amplitude segment of the curve might be a significant part of the curve. Thus, it would be underweighted and underemphasized in the curve fitting routine; a large amplitude segment that may not be of significance would be overweighted and overemphasized. Thus, the coefficients would be out of proportion to the significance of the curve. Therefore, all except the zero weights were eliminated.

G. TIME SCALING

The initial interval between data points was arbitrarily chosen to be one in the curve fitting routine. The resulting coefficients were out of proportion in that the low degree coefficients were many orders of magnitude larger than the high degree coefficients. In the comparison of coefficients of supposedly similar curves, the high order coefficients are far more important than the low order coefficients. Therefore, it was necessary to choose a more appropriate interval that would decrease the relative magnitudes of the coefficients.

The interval size is inversely proportional to the number of data points being used. The use of 200 sample points requires an interval of 0.1 units, whereas the use of 100 points requires an interval size of 0.2 units. This size requirement is based on the present state of the program.

H. SECOND DEGREE SMOOTHING

Requiring a polynomial to fit a curve with many relative maximums and minimums, many of which occur within a very short distance, causes the coefficients to inaccurately represent the envelope of the curve. By eliminating the minimum points, and keeping only the maximum points, a second degree smoothing was effected. A copy of the program segment used to accomplish this can be found at the end of the computer program section.

This method was discarded under the current program configuration because it eliminated not only unimportant segments of the curve, but it also under certain circumstances eliminated salient features of the curve.

I. DEGREE OF POLYNOMIAL FIT

In looking at the brush recordings of some of the words used, it is difficult to determine just what degree of polynomial fit is necessary to get an accurate representation of the curve in terms of coefficients. At first, a twentieth degree fit was used under the assumption that the larger the degree of the polynomial the better the fit. After plotting some of the resultant curves, it became obvious that although a twentieth degree fit was appropriate for some of the curves, it was too great a degree of fit for others because minor variations in the curve were emphasized. A tenth degree fit was then tried in order to give a better average result for all of the curves. This, too, was inappropriate in that it was too small a degree of fit. The present program performs a fifteenth degree fit for all curves.

VI. SUMMARY

Although the current system does not recognize speech, some combination of the present program and the recommendations made may lead to a speech recognizer. Two hardware limitations were encountered; it was impossible to construct eight high resolution filters on the CI 5000; and there was insufficient direct access core storage in the XDS 9300. Consequently, low resolution filters and a small sample size had to be used. One system software limitation was encountered; the data transfer subroutine, ADL, was found to be too slow, thus prohibiting high frequency sampling.

Based on the initial experimentation, and the results obtained thus far, it is possible that at least one significant speech parameter is being overlooked. Although frequency and formant analysis may be necessary, they are not sufficient for a generalized speech recognizer.

Each word and sound investigated contained a basic wave shape, but due to pronunciation differences, the shape was altered sufficiently that coefficient correlation was not effective. The extracting of distinctive portions of the curve that remain the same from trial to trial should lead to a greater degree of correlation.

VII. RECOMMENDATIONS

A. CURVE AVERAGING

Instead of comparing coefficients per se, an averaging of input data points from trial to trial and a study of the resulting coefficients, appears to be a promising approach to the problem of speech recognition using the previously described system. This would entail using overlaying techniques in the XDS-9300 system.

The main problem associated with this approach is one of timing; the beginning and end of the curves must coincide to be averaged.

B. TIME NORMALIZATION

The timing problem just mentioned in the previous section bears rectification immaterial of what other future changes are made to the program. A curve that is stretched over a longer distance bears little resemblance to the unstretched curve coefficient-wise. For this reason, any future polynomial curve fitting approach must take into account the problem of sound duration.

C. SEGMENTED CURVE FITTING

Throughout the experimentation, it was noticed that although one particular curve did not totally match another, there were large segments of the curves that matched quite well, especially in the latter segments. Thus, instead of one set of coefficients to represent an audio curve, there might be several representing various curve segments. Again, time normalization must be considered.

D. HARDWARE CHANGES

The bandpass filters used were relatively low resolution due to hardware limitations imposed by the CI-5000. In order to have better filters, it would be necessary to construct them from component parts. There is strong evidence that this would help to eliminate the harmonics of voiced audio signals, which cause random variance at different frequency ranges dependent upon the speaker.

E. SECOND DEGREE ANALOG SMOOTHING

Although digital second degree smoothing was found to be of no practical value, this does not mean that a second degree analog smoothing circuit would react in the same manner. Implementing this feature could help to alleviate minor differences in audio curves. Thus, a closer coefficient correlation could be effected.

F. INPUT DATA CORRELATION

To this point, all recommendations have concerned themselves in some manner with coefficient correlation. Given a time normalized curve, it might be interesting to attempt data point correlation of some form. As was pointed out in the section recommending segmented curve fitting, there were often parts of the audio curves that compared quite favorably. By looking only at the associated data points, an interesting type of correlation might be accomplished.

G. ORTHOGONAL COEFFICIENT CORRELATION

The current program outputs coefficients of the form B_i, as described in section II. However, each B_i is dependent upon all of the orthogonal coefficients, C_j. The equation is of the form: $B_i = K \sum_j C_j \; O_j(x)$

where the O_j (x) are orthogonal polynomials. It is obvious from this that a change in only one C_j will affect every B_i. Therefore, results could perhaps be attained by investigating the orthogonal coefficients.

```
C
C     LEAST SQUARES POLYNOMIAL CURVE FITTING
C
C     GIVEN ARRAY X AND ARRAY F2, WHERE F2(I) IS THE OBSERVED DEPENDENT
C     VARIABLE AND X(I) IS THE OBSERVED INDEPENDENT VARIABLE, THE
C     POLYNOMIAL Y=B(U)+B(2)*X+...+B(K+1)*X**K IS FITTED FOR ALL DEGREES
C     K,1.LE.K.LE.K(MAX).
C
C     M - NUMBER OF DATA POINTS, M.LE.200
C
C     KM - THE VALUE OF KM IS THE DEGREE OF FIT TO BE CONSIDERED.
C
C     X - THE ARRAY OF OBSERVED INDEPENDENT VARIABLES.
C
C     F2 - THE ARRAY OF OBSERVED DEPENDENT VARIABLES.
C
C     WI - THE ARRAY OF WEIGHTS.
C
C     Y - THE OUTPUT ARRAY OF ESTIMATED DEPENDENT VARIABLES.
C
C     DELY - THE OUTPUT ARRAY OF THE DIFFERENCES BETWEEN F2 AND Y.
C
C     B - THE OUTPUT ARRAY OF THE COEFFICIENTS OF THE POWERS OF X.
C
      NAMELIST M,JK,XX
      DIMENSION A(21,21),BUF(2000),X(400),F2(400),WI(400),Y(400),
     1B(11),F(400),P(400),PM(400),DELY(400),T(400),W(400)
C
C     ANALOG TO DIGITAL CONNECTING STATEMENT
C
      CONNECT(52,AD(BUF,I1,MM,N,40S))
C
C     INPUT DESIRED SAMPLE SIZE.........NOT TO EXCEED 200
C
      OUTPUT(101) 'FORMAT FOR INPUT IS....  M= VALUE * CR'
```

```
      INPUT(101)
C
C     INPUT DESIRED FILTER FILTER NUMBER FOR PRINTOUT OF INPUT DATA
C
      OUTPUT(101) 'FORMAT FOR INPUT IS....  JK= VALUE * CR'
      INPUT (101)
C
C     INPUT DESIRED INTERVAL SIZE
C
      OUTPUT(101) '       XX= VALUE * CR    '
      INPUT(101)
      N=0
      MM=M * 8
      X(1)=0.0
      Y(1)=0.0
      KM=15
      DO 2 I=2,M
    2 X(I) = X(I-1) + XX
C
C     ZERO OUT DATA STORAGE AREA
C
 9999 DO 11 I=1,MM
      BUF(I)=0.0
   11 CONTINUE
C
C     TEST FOR ANALOG GENERATED INTERRUPT
C
    7 IF(TEST(1).GT.0)GO TO 7
      I1=1
C
C     ENABLE HARDWARE INTERRUPT
C
      CALL ENABLE
C
C     WAITING LOOP THAT CYCLES UNTIL ANALOG INTERRUPT OCCURS
```

```
C
S   12 NOP
S      PRX 128,0
C
C      DISABLE HARDWARE INTERRUPT DUE TO COMPLETION OF DATA INPUT
C
    40 CALL DISABLE
       WRITE(6,2020)
 2020 FORMAT('1')
C
C      NOISE ELIMINATION AND SCALING LOOP
C
       DO 47 I=1,MM
       J=IFIX(BUF(I)*1000.)
    47 BUF(I)=FLOAT(J)/100.0
C
C      TEST TO SEE IF INPUT DATA PRINTOUT IS DESIRED
C
       IF(SENSE SWITCH 1)48,49
    48 CONTINUE
       WRITE(6,1234)(BUF(I), I=JK,MM,8)
 1234 FORMAT(8F15.6)
       WRITE(6,2020)
    49 CONTINUE
C
C      BEGIN CURVE FITTING ROUTINE
C
       KOUNT=7
    50 CONTINUE
       DO 3 I=1,M
       JJ=8*I-KOUNT
C
C      PASS DATA FROM BUFFER TO ARRAY OF OBSERVED DEPENDENT VARIABLES.
C
       F2(I)=BUF(JJ)
```

```
C
C     SET WEIGHTS TO EITHER ZERO OR ONE.
C
      WI(I)=1.0
    3 IF(F2(I).EQ.0.0)WI(I)=0.0
      DO 1 I=1,21
      DO 1 J=1,21
    1 A(I,J)=0.0
      A(1,1)=1.0
      A(2,2)=1.0
      FM=0.0
      DO 5 I=1,M
C
C     SUM THE WEIGHTS.
C
    5 FM=FM+WI(I)
C
C     CALCULATE RECIPROCAL OF THE SUM OF WEIGHTS.
C
      FMR=1.0/FM
      FBAR=0.0
      XBAR=0.0
      DO 10 I=1,M
      W(I)=WI(I)*FMR
      PM(I)=SQRT(W(I))
      F(I)=F2(I)*PM(I)
C
C     CALCULATE THE WEIGHTED MEANS.
C
      FBAR=FBAR+F(I)*PM(I)
   10 XBAR=XBAR+X(I)*W(I)
      T(1)=FBAR
      A(2,1)=-XBAR
      PXF=0.0
      PXP=0.0
```

```
      DO 20 I=1,M
      P(I)=(X(I)-XBAR)*PM(I)
      PXF=PXF+P(I)*F(I)
   20 PXP=PXP+P(I)*P(I)
      T(2)=PXF/PXP
      PMXPM=1.0
      KMP=KM+1
C
C     FIRST VALUES FOR B(1) AND B(2) NECESSARY FOR ITERATION.
C
      B(1)=T(1)*A(1,1)+T(2)*A(2,1)
      B(2)=T(2)*A(2,2)
      SUMSQ=0.0
C
C     BEGIN COMPUTING COEFFICIENTS ITERATIVELY.
C
      DO 1000 K=2,KMP
      KM1=K-1
      KM2=K-2
      IF(KM2.EQ.0)GO TO 1000
      XPXP=0.0
      XPXPM=0.0
      B(K)=0.0
      DO 51 J=1,M
      XP=X(J)*P(J)
      XPXP=XPXP+XP*P(J)
   51 XPXPM=XPXPM+XP*PM(J)
      ALPHA=XPXP/PXP
      BETA=XPXPM/PMXPM
      PPXF=0.0
      PPXPP=0.0
      DO 100 I=1,M
      PT=P(I)
      P(I)=(X(I)-ALPHA)*P(I)-BETA*PM(I)
      PPXF=PPXF+P(I)*F(I)
```

```
      PPXPP=PPXPP+P(I)*P(I)
  100 PM(I)=PT
      T(K)=PPXF/PPXPP
      PMXPM=PXP
      PXP=PPXPP
      A(K,1)=-ALPHA*A(KM1,1)-BETA*A(KM2,1)
      A(K,KM1)=A(KM1,KM2)-A(KM1,KM1)*ALPHA
      A(K,K)=1.0
      IF(K.LE.3)GO TO 150
      DO 120 I=2,KM2
  120 A(K,I)=A(KM1,I-1)-ALPHA*A(KM1,I)-BETA*A(KM2,I)
  150 DO 160 I=1,K
  160 F(I)=B(I)+T(K)*A(K,I)
C
C     BEGIN COMPUTATION OF DEPENDANT VARIABLES BASED ON COMPUTED COEFFIC
C
      DO 222 I=2,M
      SQ=B(K)
      DO 223 IQ=1,KM1
      KMIQ=K-IQ
  223 SQ=X(I)*SQ+B(KMIQ)
      Y(I)=SQ
      DELY(I)=Y(I)-F2(I)
  222 SUMSQ=SUMSQ+DELY(I)*DELY(I)
C
C     END OF DEPENDANT VARIABLE COMPUTATION LOOP
C
 1000 CONTINUE
      NUM=8-KOUNT
      DO 556 I=1,2
      WRITE(6,555)
  555 FORMAT(' ')
  556 CONTINUE
C
C     COEFFICIENT NORMALIZATION LOOP
```

```
C
      DO 444 I=1,16
  444 B(I) = .0001 * B(I) / B(13)
C
C     BEGIN PRINTOUT OF CALCULATED COEFFICIENTS
C
      WRITE(6,333)NUM
  333 FORMAT(' COEFFICIENTS FOR FILTER',I2,/)
      WRITE(6,9520)((I,B(I)),I=1,KMP)
 9520 FORMAT(6('  B(',I2,')=',F12.5))
C
C     TEST TO SEE IF PRINTOUT OF DEPENDANT VARIABLES IS DESIRED FOR PRES
C     FILTER
C
      IF(SENSE SWITCH 2)264,265
  264 CONTINUE
      WRITE(6,555)
      WRITE(6,1111)(Y(I), I=1,M)
 1111 FORMAT(8F15.6)
      WRITE(6,2020)
  265 KOUNT=KOUNT-1
C
C     TEST TO SEE IF DATA FROM ALL EIGHT FILTERS HAVE HAD
C     COEFFICIENTS CALCULATED
C
C     IF TEST IS TRUE GO TO 9999, ZERO OUT THE DATA AREA AND WAIT
C     FOR ANOTHER ANALOG GENERATED INTERRUPT
C
C     IF TEST IS FALSE GO TO 50, AND CALCULATE THE COEFFICIENTS
C     FOR THE NEXT FILTER'S DATA
C
      IF(KOUNT.LT.0)GO TO 9999
      GO TO 50
      END
```

```
      SUBROUTINE AD(BUF,I1,MM,N,LOC)
      DIMENSION BUF(2000)
C
C     ONE DATA POINT FROM EACH FILTER IS ENTERED INTO BUFFER
C
      CALL ADL(0,BUF(I1),BUF(I1+1),BUF(I1+2),BUF(I1+3),BUF(I1+4),
     1BUF(I1+5),BUF(I1+6),BUF(I1+7))
C
C     INCREMENT THE BUFFER INDEX
C
      I1=I1+8
C
C     TEST TO SEE IF ALL DATA IS IN THE BUFFER.   IF TRUE RETURN TO STAT
C     NUMBER 40 IN MAIN PROGRAM, DISABLE INTERRUPT AND BEGIN THE ANALYSI
C
C     IF FALSE, RETURN TO JUMPJ LOOP IN MAIN PROGRAM AND AWAIT INTERRUPT
C     FOR NEXT SET OF SAMPLE POINTS
C
      IF(I1.GT.MM) RETURN LOC
      RETURN
      END
```

```
C      SECOND DEGREE SMOOTHING - INSERT IMMEDIATELY AFTER STATEMENT #2
C
C
C
C      ISW - SWITCH THAT DENOTES SIGN OF SLOPE
C
C      NPOINT - NUMBER OF DATA POINTS EXCLUDING TRAILING ZEROS
C
       ISW=1
     6 I=J-8-KOUNT
     8 IF(BUF(I+8).GE.BUF(J))GO TO 9
       I=I+8
       IF(ISW.EQ.0),BUF(J)>BUF(I);GO TO 13
       J=J+8
       BUF(J)=BUF(I)
       ISW=0
    13 IF(I.GT.MM)GO TO 13
       GO TO 8
     9 I=I+8
       BUF(J)=BUF(I)
       ISW=1
       GO TO 13
    14 L=J+8
       DO 15 I=LBOUN,MM,8
    15 BUF(I)=0.0
       NPOINT=(J+KOUNT)/8
C
C      END OF DIGITAL SMOOTHING SEGMENT
```

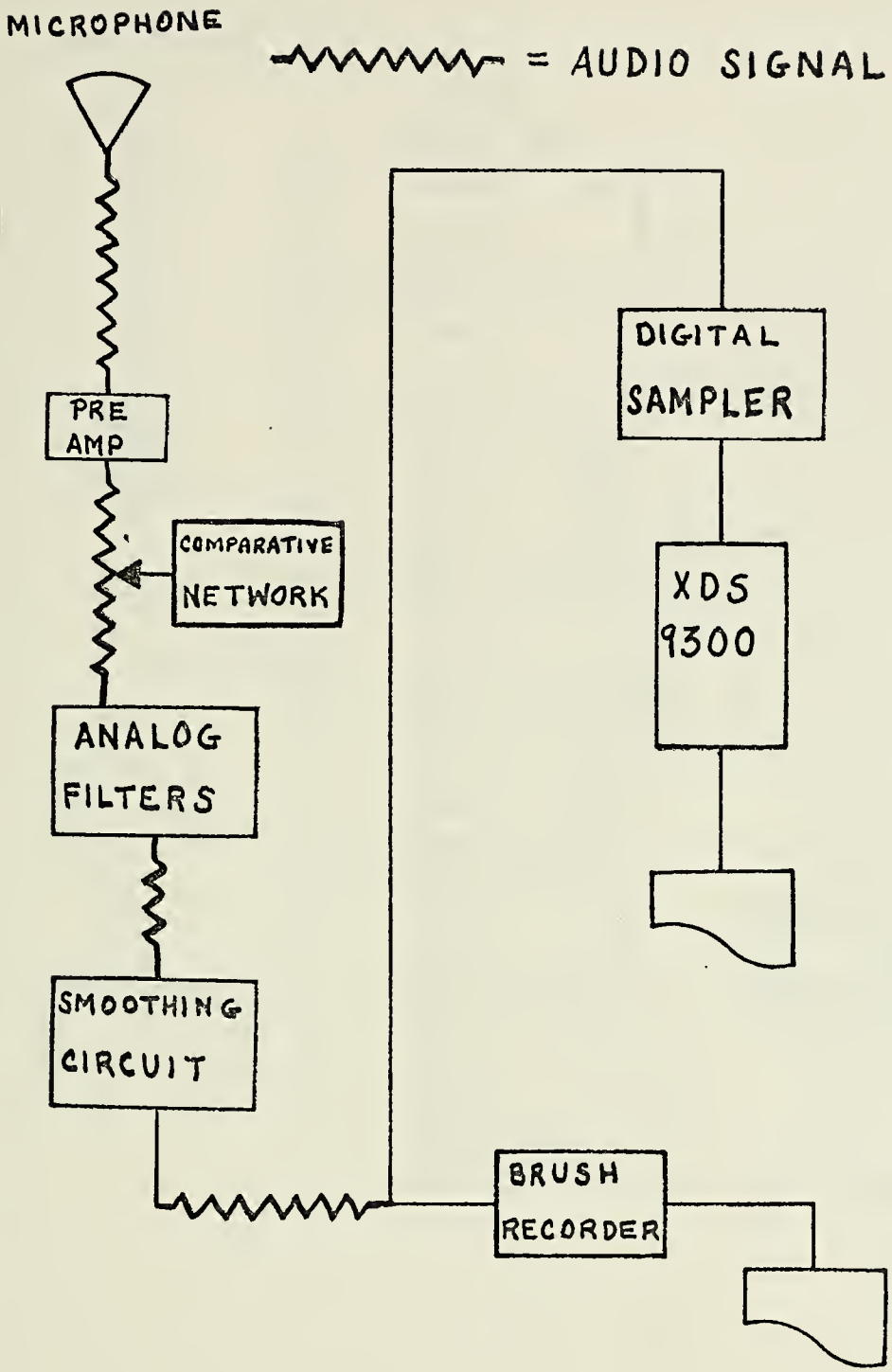
MICROPHONE
= AUDIO SIGNAL
PRE AMP
COMPARATIVE NETWORK
ANALOG FILTERS
SMOOTHING CIRCUIT
DIGITAL SAMPLER
XDS 9300
BRUSH RECORDER

FIGURE 1

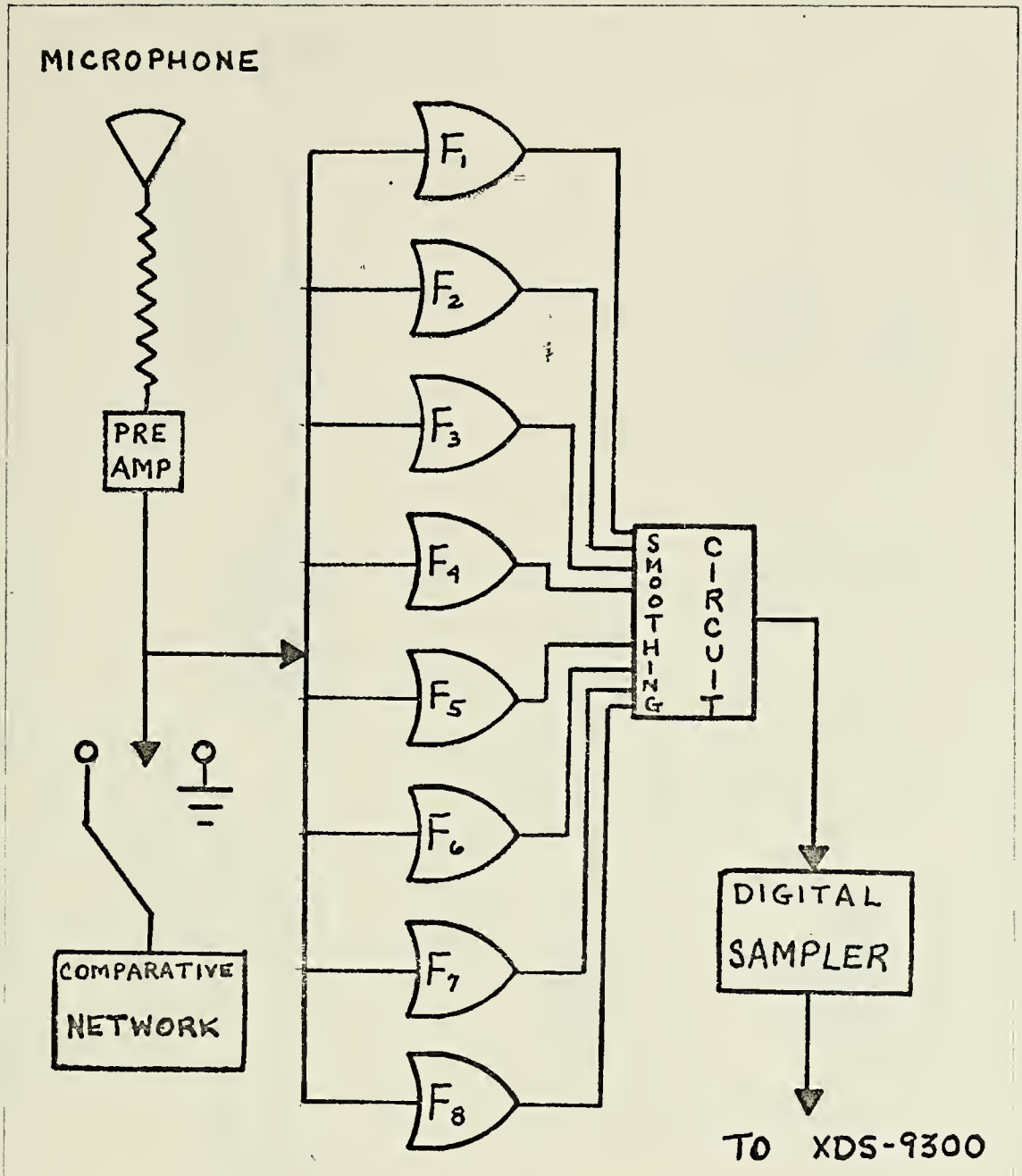
MICROPHONE
PRE AMP
F1
F2
F3
F4
F5
F6
F7
F8
SMOOTHING CIRCUIT
COMPARATIVE NETWORK
DIGITAL SAMPLER
TO XDS-9300

FIGURE 2

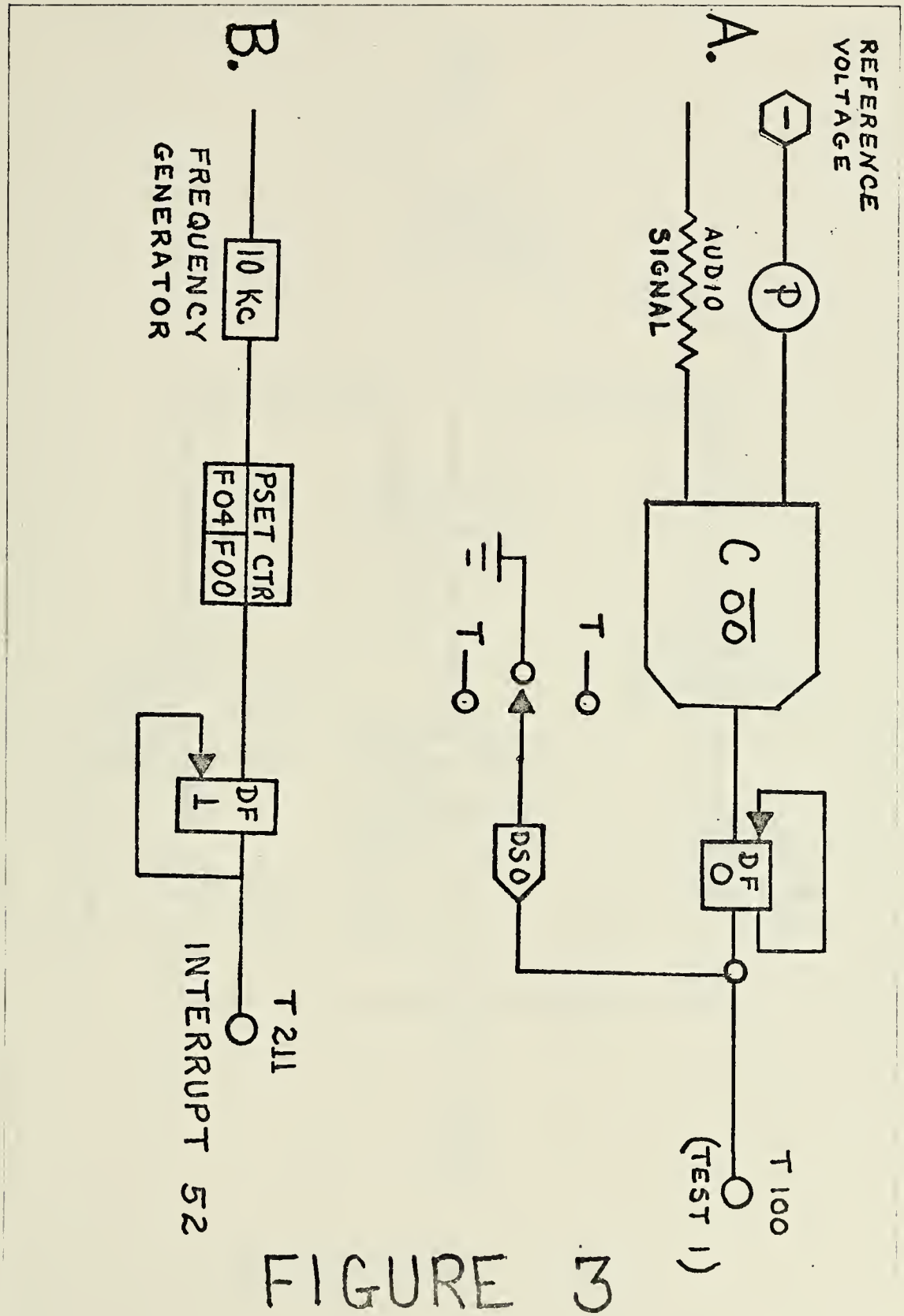
A.
REFERENCE VOLTAGE
AUDIO SIGNAL
C 00
DF 0
T 100
(TEST 1)
T
T
DSO
B.
FREQUENCY GENERATOR
10 Kc
PSET CTR
F04 F00
DF 1
T 211
INTERRUPT 52

FIGURE 3

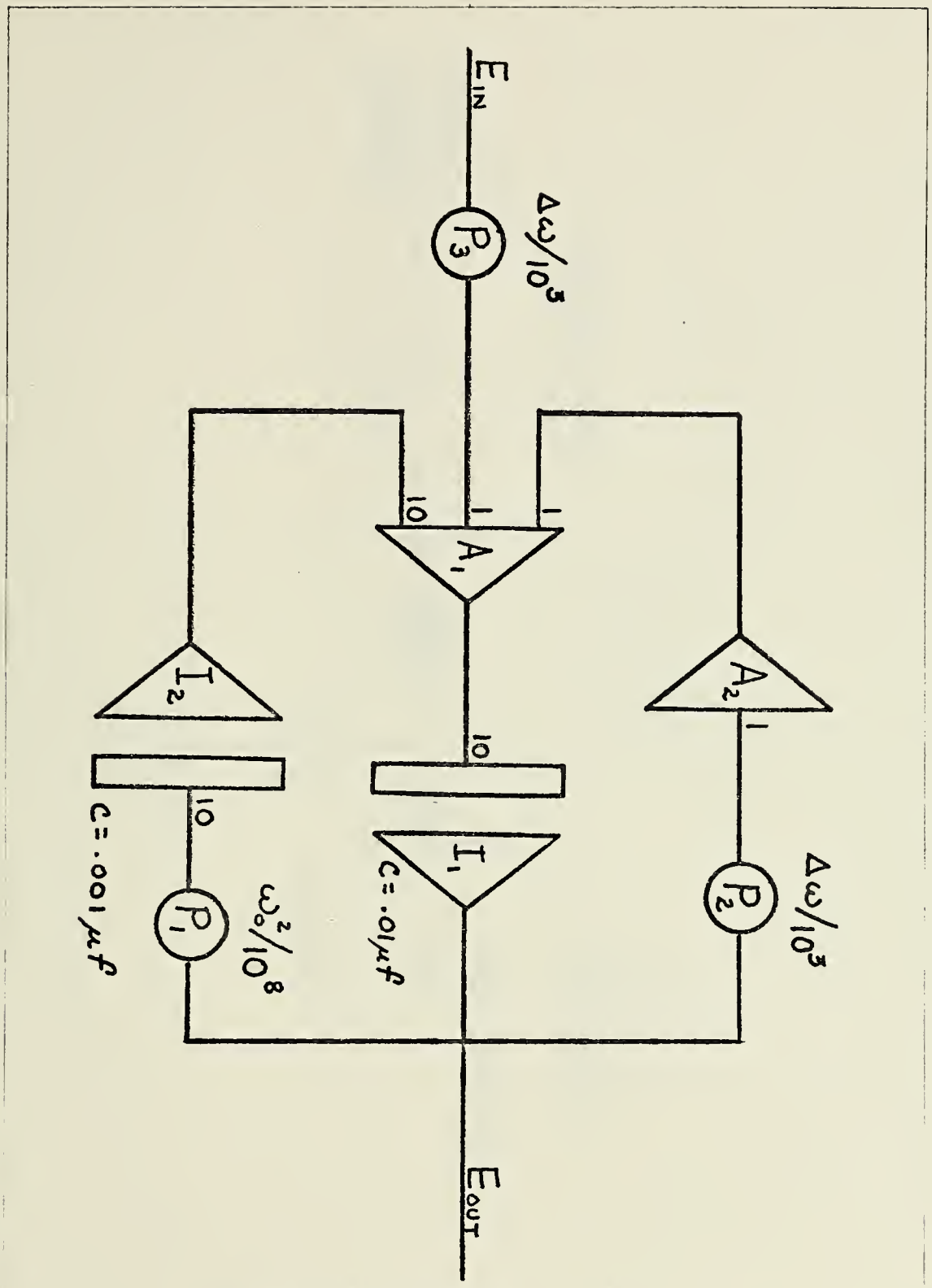
E_{IN}
P_3
$\Delta\omega/10^5$
10
1
1
A_1
I_2
A_2
10
1
10
$C = .001\,\mu f$
I_1
$C = .01\,\mu f$
P_1
$\omega_o^2/10^8$
P_2
$\Delta\omega/10^3$
E_{OUT}

FIGURE 4

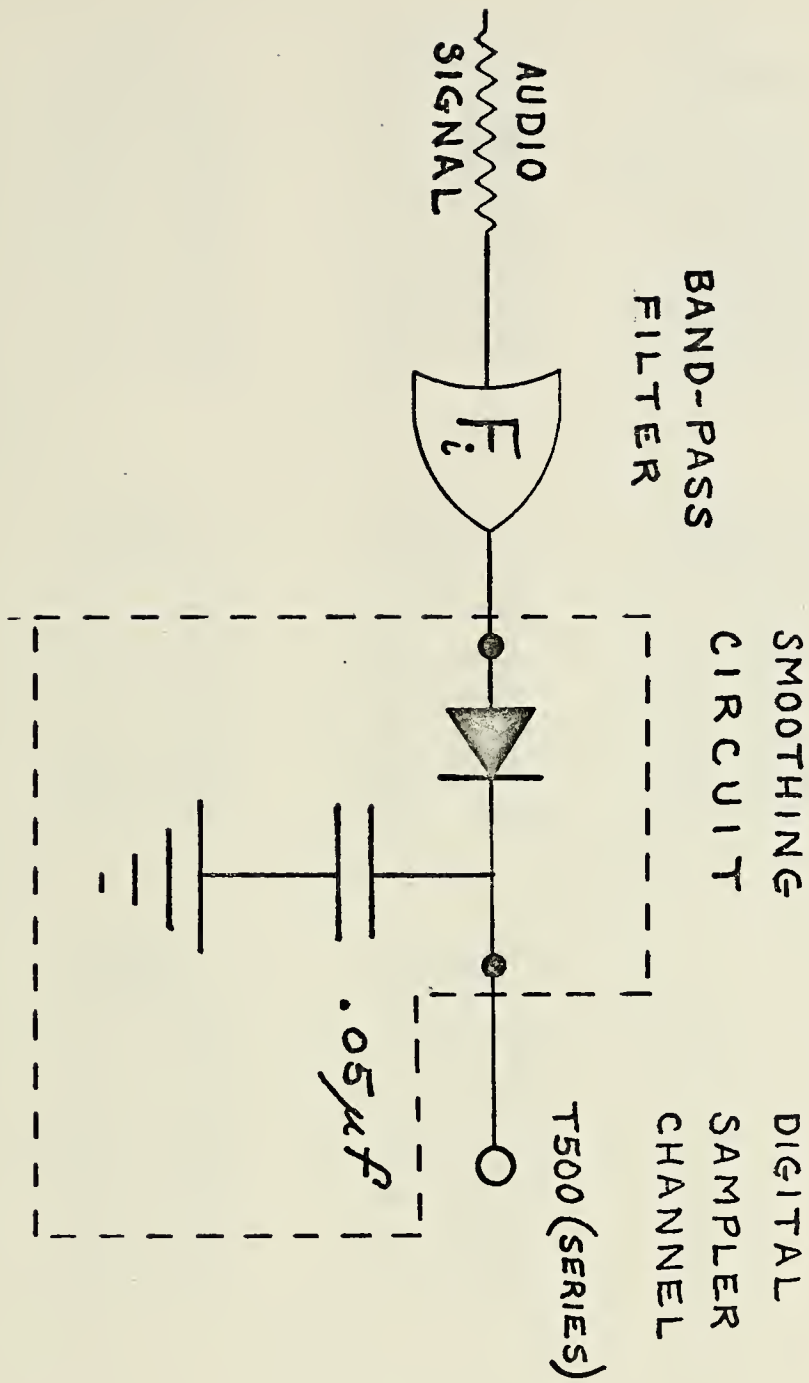
AUDIO SIGNAL
BAND-PASS FILTER
F_i
SMOOTHING CIRCUIT
.05 μf
T500 (SERIES)
DIGITAL SAMPLER CHANNEL

FIGURE 5

FILTER 1	FILTER 2	FILTER 3
I1 = A001	I1 = A005	I1 = A011
I2 = A063	I2 = A007	I2 = A013
A1 = A000	A1 = A010	A1 = A014
A2 = A002	A2 = A006	A2 = A016
P1 = P001	P1 = P005	P1 = P011
P2 = P000	P2 = P004	P2 = P012
P3 = P002	P3 = P006	P3 = P013

FILTER 4	FILTER 5	FILTER 6
I1 = A015	I1 = A031	I1 = A041
I2 = A017	I2 = A033	I2 = A037
A1 = A022	A1 = A026	A1 = A034
A2 = A024	A2 = A030	A2 = A036
P1 = P015	P1 = P031	P1 = P037
P2 = P016	P2 = P027	P2 = P036
P3 = P021	P3 = P026	P3 = P034

FILTER 7	FILTER 8
I1 = A045	I1 = A053
I2 = A047	I2 = A055
A1 = A042	A1 = A050
A2 = A044	A2 = A052
P1 = P045	P1 = P053
P2 = P044	P2 = P052
P3 = P042	P3 = P050

TABLE 1

POTENTIOMETER SETTINGS

P000	.1885	
P001	.1006	
P002	.1885	
P004	.1885	
P005	.1264	
P006	.1885	
P010	.1641	
P012	.1885	
P013	.1885	
P015	.2015	
P016	.1885	
P021	.1885	
P026	.1885	
P027	.1885	
P031	.2429	
P034	.1885	
P036	.1885	
P037	.2883	
P042	.1885	
P044	.1885	
P045	.3376	
P050	.1885	
P052	.1885	
P053	.3904	
P406	.020	biasing pot

	LOWER 3 db LEVEL	CENTER FREQ	UPPER 3 db LEVEL	
FILTER 1	490	505	520	
FILTER 2	560	575	590	
FILTER 3	630	645	660	
FILTER 4	700	715	730	all frequencies
FILTER 5	770	785	800	in hertz (Hz)
FILTER 6	840	855	870	
FILTER 7	910	925	940	
FILTER 8	980	995	1010	

TABLE 2

REFERENCES

1. Defense Documentation Center Report CS49, An Approach to Computer Speech recognition by Direct Analysis of the Speech Wave by D. R. Reddy, pp 1-143, 1 September 1966.

2. Reddy, D. R. and Vicens, P. J., "A Procedure for the Segmentation of Connected Speech," Journal of the Audio Engineering Society, v.16, pp 404-411, October 1968.

3. Bobrow, D. G. and Klatt, D. h., "A Limited Speech Recognition System," AFIPS Conference Proceedings, v.33, pp 305-318, 1968

4. Daniloff, R. G., Shriner, T. H. and Zemlin, W. R., "Intelligibility of Vowels Altered in Duration and Frequency," Journal of the Acoustical Society of America, v.44, pp 700-707, September 1968.

5. Seo, H., "Speech Compression," Dissertation Abstracts, v.288, p 3713, March 1968.

6. United States Government Research and Development Reports, Reprocessing for Speech Analysis by P. Vicens, p 129, 10 January 1969.

7. Walker, D., Personal Interview, Stanford Research Institute, 15 February 1972.

BIBLIOGRAPHY

Allen, J., "Man-to-Machine Communications by Speech Part II: Synthesis of Prosodic Features of Speech by Rule," Spring Joint Computer Conference, p. 339-343, 1968.

Bell, C. G., and others, "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," Journal of The Acoustical Society of America, v. 33, p. 1725-1736, December, 1961.

Hill, D. R., Pattern Recognition, p. 199-226, American Elsevier, 1966.

Massachusetts Institute of Technology Research Laboratory of Electronics Report 395, The Recognition of Speech by Machine, by G. W. Hughes, p. 1-60, 1 May 1961.

Kurland, M. and Papson, T. P., "Analog Computer Simulation of Linear Modulation Systems," Analog/Hybrid Computer Educational Society Transactions, v. III, p. 13, January 1971.

Lavington, S. H. and Rosenthal, L. E., "Some Facilities for Speech Processing by Computer," Computer Journal, v. 9, p. 330-339, February 1967.

Lee, F. F., "Machine-to-Man Communications by Speech Part 1: Generation of Segmented Phonemes from Text," Spring Joint Computer Conference, p. 333-338, 1968.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	LTJG Robert C. Bolles, Code 53Bq Department of Mathematics Naval Postgraduate School Monterey, California 93940	1
4.	LT Frederick M. Stubbs 373-E Bergin Drive Monterey, California 93940	1
5.	LT John Paul Hydinger 373-E Bergin Drive Monterey, California 93940	1

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

An Investigation Into Machine Recognition of Vowel-like Sounds and Their Allophones in One Syllable Words

4. DESCRIPTIVE NOTES *(Type of report and, inclusive dates)*
Master's Thesis; June 1972

5. AUTHOR(S) *(First name, middle initial, last name)*
John Paul Hydinger
Frederick Michael Stubbs

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1972	50	7

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

The goal was to recognize sustained vowel-like sounds and their allophones in one syllable words. A bank of filters and a digital sampler provided a data base for a polynomial curve fitting routine. The frequency range under investigation was 500-1000 Hz. A COMCOR CI 5000 analog computer and an XDS 9300 digital computer were used. Although coefficient correlation was ineffective, several recommendations for system improvement are made.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

A-31408

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
peech Recognition						
attern Recognition						
owel Analysis						
ord Analysis